

Number of clusters, deconvolution and classical problem of moments

Lev B. Klebanov*, Zeev Volkovich†

Abstract

In the paper there is given a connection between one special case of cluster analysis, deconvolution problem, and classical moment problem. Namely, the methods used there are applied to solve deconvolution problem for the case of one known distribution and another one concentrated in unknown finite number of points. These results can be applied to estimate a number of clusters for the case of scale or location mixture of identical distributions.

keywords: number of clusters, deconvolution problem, classical moment problem, scale and location mixtures.

1 Introduction

We study deconvolution problem for two cases. The first variant is to estimate one of the multiplicative convolution components when another component is known. The convolution itself is observable, and unknown component is concentrated at finite number of points. The second variant concerns the case of additive convolution.

Let us explain our problem and method on one simple example. Suppose that Z is a random variable having standard normal distribution. Let Y be a random variable, independent of Z and taking k unknown positive values $\sigma_1, \sigma_2, \dots, \sigma_k$ with probabilities $p_j = \mathbb{P}\{Y = \sigma_j\}$, $j = 1, \dots, k$. The

*Department of Probability and Mathematical Statistics, MFF, Charles University, Prague 18675, Czech Republic.

†Software Engineering Department, ORT Braude College of Engineering, Karmiel 21982, Israel.

number k is supposed to be fixed, but unknown. Denote $X = Y \cdot Z$, and suppose that we have n observations of X , that is n independent identically distributed (i.i.d.) with X random variables X_1, \dots, X_n . Our aim is to estimate the number k and the parameters σ_j , $j = 1, \dots, k$ on the base of observations $\{X_i, i = 1, \dots, n\}$. It is possible to interpret the distribution of X as normal laws scale mixture. Each normal distribution with variance σ_j^2 may be considered as a cluster. The number of clusters is k , and the weight of j th cluster is p_j . We have to estimate numbers k and p_j , ($j = 1, \dots, k$). In view of independence of Z and Y , the moments $\mu_j(X) = \mathbb{E}X^j$, ($j = 1, 2, \dots$) are products of corresponding moments $\mu_j(Z)$ and $\mu_j(Y)$:

$$\mu_j(X) = \mu_j(Z)\mu_j(Y), \quad j = 1, 2, \dots$$

Because the distribution of Z is known, we may estimate the moments $\mu_j(Y)$ using empirical moments instead of $\mu_j(X)$. Without loss of generality, we may change random variable Y by another variate (denoted by Y again) which has symmetric distribution taking the values $\pm\sigma_j$ with probabilities $p_j/2$ each, so that $\mu_{2*j-1}(Y) = 0$ for $j = 1, 2, \dots$. Now we may apply the theory for classical problem of moments (see, for example, [1]). The condition of positiveness of thruncated sequence of moments gives us an estimate for k , while the parameters p_j may be estimated on the base of corresponding orthogonal polynomials.

The problem mentioned above may be considered as deconvolution problem. To see this it is sufficient to pass from random variables X, Y, Z to logarithms of their absolute values. It allows to pass from multiplicative mixture to additive one. The problem takes now the folloowing form. It is necessary to estimate one component of its convolution with known distribution basing on observation of the convolution itself. It is so-called deconvolution problem.

2 Procedure for a solution of deconvolution problem in the case of scale mixture

Here we give detailed procedure of deconvolution problem for the case of scale mixtures. We use notations of Section 1, but the distribution of Z is not supposed to be Gaussian. However, we will impose some restrictions on the distribution.

Assumption A. The distribution of random variable Z possesses of all moments and can be recovered from them in unique way.

Assumption B. All moments of random variable Z are non-zero.

Let $X = Z \cdot Y$ be a product of independent random variables, where the distribution of Z is known and satisfies Assumptions A and B. Y is a random variable, independent of Z and taking k unknown positive values $\sigma_1, \sigma_2, \dots, \sigma_k$ with probabilities $p_j = \mathbb{P}\{Y = \sigma_j\}, j = 1, \dots, k$. The number k is supposed to be fixed, but unknown. Basing on a random sample X_1, \dots, X_n from the distribution of X we have to estimate k and $\sigma_1, \dots, \sigma_k$.

Suppose that both **Assumptions A and B** hold. Denote by $m_j(n) = \frac{1}{n} \sum_{i=1}^n X_i^j, j = 1, 2, \dots$ - empirical moments of X . Then, the value

$$\mu_j^*(n) = m_j / \mu_j(Z), \quad j = 0, 1, 2, \dots \quad (2.1)$$

may be used as consistent estimator of $\mu_j(Y)$. Because random variable Y is concentrated in finite number k of points, then

$$\mathcal{D}_s = \det \begin{pmatrix} \mu_0(Y) & \mu_1(Y) & \dots & \mu_s(Y) \\ \mu_1(Y) & \mu_2(Y) & \dots & \mu_{s+1}(Y) \\ \dots & \dots & \dots & \dots \\ \mu_s(Y) & \mu_{s+1}(Y) & \dots & \mu_{2s}(Y) \end{pmatrix} > 0, \quad (2.2)$$

for all $s < k$, and determinant (2.2) equals to 0 for $s = k$ and, therefore, for all $s > k$ (see, [2]). Therefore, we propose to use the following procedure for estimation of the number k . Consider determinants

$$\mathcal{D}_s^* = \det \begin{pmatrix} 1 & \mu_1^*(n) & \dots & \mu_s^*(n) \\ \mu_1^*(n) & \mu_2^*(n) & \dots & \mu_{s+1}^*(n) \\ \dots & \dots & \dots & \dots \\ \mu_s^*(n) & \mu_{s+1}^*(n) & \dots & \mu_{2s}^*(n) \end{pmatrix}, \quad (2.3)$$

for the first value s for which $\mathcal{D}_s^* \leq 0$. This value denote by $k^* = k^*(n)$, and use it as an estimator for k . $k^*(n)$ is a consistent estimator of k because (2.1) are consistent estimators for the moments of random variable Y . In other words, $k^*(n) \rightarrow k$ as $n \rightarrow \infty$ in probability.

Let us make some remarks about this procedure:

1. Estimators $\mu_j^*(n)$ of $\mu_j(Y)$ are not moments themselves. Therefore, the determinants \mathcal{D}_s^* are not obligated to be non-negative for all s .

2. If the number k is not small, then the order of first zero determinant (2.2) is $2k$, which is also not small, and its calculation is not simple. We need high accuracy of calculations and, therefore, high precision of moments estimation, and, consequently, large number of needed observations n .
3. It is known, that the accuracy of estimators of the moments of high order is rather low. Therefore, we have one more argument, that the number of observations n has to be very large for not small values of the components number k .

The problem of estimating the number of clusters is very interesting by itself, and therefore we provide simulation study of the quality of estimator k^* . The problem of estimating $\sigma_1, \sigma_2, \dots, \sigma_k$ will be considered later.

We simulated samples of different sample size from mixtures of exponential distributions. On their basis we estimated the number k of components in the mixtures. The results are given below.

- Simulation of samples from mixture of $k = 2$ exponential distributions with scale parameters $\lambda_1 = 1$ and $\lambda_2 = 1/3$ (mean values are 1 and 3 correspondingly) with equal weights. Sample size used $n = 100$. There were 5000 simulations. The estimator $k^* = k^*(100)$ had the following parameters: mean value $\mathbb{E}(k^*) = 1.951$; median $me(k^*) = 2$; standard deviation $\sigma(k^*) = 0.236$; percent of correct defined values $pr(k^*) = 94.18\%$. The results here seem to be rather good.
- Simulation of samples from mixture of $k = 3$ exponential distributions with scale parameters $\lambda_1 = 1$, $\lambda_2 = 3$ and $\lambda_3 = 5$ (mean values are 1, 3 and 5) with equal weights. Sample size used $n = 100$. There were 5000 simulations. The estimator $k^* = k^*(100)$ had the following parameters: mean value $\mathbb{E}(k^*) = 1.98$; median $me(k^*) = 2$; standard deviation $\sigma(k^*) = 0.182$; percent of correctly defined values $pr(k^*) = 1.14\%$. The results seem to be very bad. However, it is clear, that the sample size is too small for such problem.
- Now the same situation as before, but with different sample sizes. We will see that: the mean value of k^* becomes closer to $k = 3$ with growing n , and percent of correctly defined values of k grows with growing n , too.

a) $n = 1,000$; $\mathbb{E}(k^*) = 2.288$; $me(k^*) = 2$; $\sigma(k^*) = 0.455$; $pr(k^*) = 29.22\%$.

b) $n = 2,000$; $\mathbb{E}(k^*) = 2.401$; $me(k^*) = 2$; $\sigma(k^*) = 0.49$; $pr(k^*) = 40.12\%$.

c) $n = 3,000$; $\mathbb{E}(k^*) = 2.456$; $me(k^*) = 2$; $\sigma(k^*) = 0.498$; $pr(k^*) = 45.58\%$.

d) $n = 4,000$; $\mathbb{E}(k^*) = 2.495$; $me(k^*) = 2$; $\sigma(k^*) = 0.50$; $pr(k^*) = 49.48\%$.

e) $n = 5,000$; $\mathbb{E}(k^*) = 2.535$; $me(k^*) = 3$; $\sigma(k^*) = 0.499$; $pr(k^*) = 53.44\%$.

f) $n = 6,000$; $\mathbb{E}(k^*) = 2.542$; $me(k^*) = 3$; $\sigma(k^*) = 0.499$; $pr(k^*) = 54.16\%$.

g) $n = 7,000$; $\mathbb{E}(k^*) = 2.576$; $me(k^*) = 3$; $\sigma(k^*) = 0.496$; $pr(k^*) = 57.36\%$.

h) $n = 8,000$; $\mathbb{E}(k^*) = 2.614$; $me(k^*) = 3$; $\sigma(k^*) = 0.489$; $pr(k^*) = 61.16\%$.

i) $n = 9,000$; $\mathbb{E}(k^*) = 2.620$; $me(k^*) = 3$; $\sigma(k^*) = 0.488$; $pr(k^*) = 61.7\%$.

j) $n = 10,000$; $\mathbb{E}(k^*) = 2.651$; $me(k^*) = 3$; $\sigma(k^*) = 0.481$; $pr(k^*) = 64.72\%$.

For $n = 30,000$ we have $\mathbb{E}(k^*) = 2.859$; $me(k^*) = 3$; $\sigma(k^*) = 0.422$; $pr(k^*) = 80.22\%$.

- Simulation of samples from mixture of $k = 4$ exponential distributions with scale parameters $\lambda_1 = 1$, $\lambda_2 = 1/3$, $\lambda_3 = 1/5$ and $\lambda_4 = 1/7$ shows that one needs from 100,000 to 1,000,000 observations to get the results similar to that for $k = 3$. We omit the exact numbers.

From all above we can see, that to make sure there exist one or two clusters, it is enough to have hundreds observations. To make sure there are three clusters we need some thousands observations. For four components there is necessarily to have hundreds of thousands or just millions of observations. The reasons for that are mentioned in 1.–3. However, there is another, more essential, reason consisting in the fact that many-components system can be often approximated rather good by three-components distributions. To

see the difference between that one needs much more observations than for small-components cases. Of course, the number of observations needed for a good agreement between estimator and reality depends of the Y distribution structure.

Let us now return to the problem of estimating the parameters $\sigma_1, \sigma_2, \dots, \sigma_k$. Define a sequence of polynomials

$$P_s^*(\lambda) = P_s^*(\lambda; n) = \frac{1}{\sqrt{\mathcal{D}_{s-1}^* \mathcal{D}_s^*}} \det \begin{pmatrix} 1 & \mu_1^*(n) & \dots & \mu_s^*(n) \\ \mu_1^*(n) & \mu_2^*(n) & \dots & \mu_{s+1}^*(n) \\ \dots & \dots & \dots & \dots \\ \mu_{s-1}^*(n) & \mu_s^*(n) & \dots & \mu_{2s-1}^*(n) \\ 1 & \lambda & \dots & \lambda^s \end{pmatrix}, \quad (2.4)$$

where \mathcal{D}_s^* is defined by (2.3). It is clear, that (2.4) gives us a consistent estimator for the following polynomial

$$P_s(\lambda) = \frac{1}{\sqrt{\mathcal{D}_{s-1} \mathcal{D}_s}} \det \begin{pmatrix} 1 & \mu_1 & \dots & \mu_s \\ \mu_1 & \mu_2 & \dots & \mu_{s+1} \\ \dots & \dots & \dots & \dots \\ \mu_{s-1} & \mu_s & \dots & \mu_{2s-1} \\ 1 & \lambda & \dots & \lambda^s \end{pmatrix}, \quad (2.5)$$

$s = 1, 2, \dots, k-1$. We cannot use this expression for $s = k$ because $\mathcal{D}_k = 0$. Therefore, we omit the multiplier $1/\sqrt{\mathcal{D}_{s-1} \mathcal{D}_s}$ for the case $s = k$. Put $P_0(\lambda) = 1$, so that the polynomials $P_s(\lambda)$ are defined for $s = 0, 1, \dots, k$. We can define $P_s^*(\lambda)$ for $s = 0$ and $s = k^*$ in a natural way. By solving equation

$$P_{k^*}(\lambda) = 0 \quad (2.6)$$

we find its roots $\lambda_1^*, \lambda_2^*, \dots, \lambda_{k^*}^*$. These values are consistent estimators for the roots $\lambda_1, \lambda_2, \dots, \lambda_k$ of polynomial $P_k(\lambda)$. It is known [1], that the roots of $P_k(\lambda)$ are the points of growth of the distribution of random variable Y . However, as has been mentioned above, the values $\mu_j^*(n)$ are not moments of any distribution, and therefore determinant \mathcal{D}_{k^*} may be negative. To finish the process of estimation of the Y distribution we need to estimate the weights σ_j at points λ_j , ($j = 1, 2, \dots, k$). It is known (see [1]), that

$$\sigma_j = \frac{1}{\sum_{i=0}^{k-1} |P_i(\lambda_j)|^2}, \quad j = 1, \dots, k. \quad (2.7)$$

From (2.7) we see that

$$\sigma_j^* = \frac{1}{\sum_{i=0}^{k^*-1} |P_i^*(\lambda_j^*)|^2}, \quad j = 1, \dots, k^* \quad (2.8)$$

are consistent estimators of σ_j .

Summarizing, we can give the complete procedure for estimating of the Y distribution under **Assumptions A and B** as follows.

1. Calculate determinants (2.3) until the first s for which $\mathcal{D}_s^* \leq 0$. This value denote by $k^* = k^*(n)$, and use it as an estimator for k .
2. Calculate polynomials (2.4) for $s = 1, \dots, k^* - 1$ and $P_0^*, P_{k^*}^*(\lambda)$.
3. By solving equation (2.6) find its roots $\lambda_1^*, \lambda_2^*, \dots, \lambda_{k^*}^*$. These roots are estimators for growth points of the Y distribution.
4. Values (2.8) give us estimators for the weights of Y distribution.
5. Finally, the estimator for Y distribution is a measure, concentrated in k^* points $\lambda_1^*, \lambda_2^*, \dots, \lambda_{k^*}^*$ and giving them weights σ_j^* , ($j = 1, \dots, k^*$).

Is it possible to avoid **Assumption B**? In general situation the answer is negative. The reason is that we cannot find corresponding moments of Y in the case when Z has some moments equal to zero. However, one can still get some information on Y distribution. For example, we can change the moments of X by its absolute moments. This will allow us to estimate consistently the distribution of its absolute value. The number of growth points for $|Y|$ is not less than a half of that for Y . The points of growth them selfs for Y are concentrated at that of $|Y|$ or at points symmetric to them. Let us note, that for the case of $|Y|$ all points of growth have to be non-negative.

3 Procedure for a solution of deconvolution problem in the case of location mixture

Here we give detailed procedure of deconvolution problem for the case of location mixtures. We use notations of Sections 1, 2. Although the case of location mixtures may be transformed to that of scale by passing to exponents

of corresponding random variables, the direct calculations appears to work better, and this is the reason for separate consideration of the location case.

So, we start with consideration of the scheme

$$X = Z + Y, \quad (3.1)$$

where X is observable random variable, Z and Y are independent random variables, the distribution of Z is known, while about that of Y distribution is known only that it is concentrate in a finite (unknown) number of points. We supposed that the **Assumption A** holds. From (3.1) it follows that

$$\mu_s(X) = \sum_{j=0}^s \binom{s}{j} \mu_j(Z) \mu_{s-j}(Y), \quad s = 1, 2, \dots$$

From here it is easy to obtain a formula for recurrent calculation of estimators for the moments of Y :

$$\mu_0^*(n) = 1, \quad \mu_s^*(n) = m_s - \sum_{j=1}^s \binom{s}{j} \mu_j(Z) \mu_{s-j}^*(n), \quad s = 1, 2, \dots \quad (3.2)$$

The rest of procedure is the same as in Section 2.

4 Simulations and Applications

Here we give some simulations and applications to analysis of stocks prices and to currency exchange rate.

4.1 Simulations

Here we give simulation results for both scale and location mixtures.

- I. *In the case of scale mixture* we simulated 5,000 samples of volume $n = 1,000$ each from mixture of two ($k = 2$) normal distributions $N(0, 1)$ and $N(0, 9)$ with equal weighs. The estimator k^* took true value 2 in 3,084 cases. Mean value of estimators for the standard deviation (equal to 1 in general population) of the first component was 0.8065. Its weight had mean value 0.4752. For the second component of the mixture we had mean value of estimators for the standard deviation

(3 in general population) equals to 2.9835. Its weight was estimated as 0.5201 in the mean. From these simulations we see that the estimator for mean of the first component is biased to the left. However, this estimator may be used as initial approximation for maximum likelihood estimation (or other) procedure.

- II. *In the case of location mixture* we simulated 5,000 samples of volume $n = 1,000$ each from mixture of two ($k = 2$) normal distributions $N(-1, 1)$ and $N(1, 1)$ with equal weights. The estimator k^* took true value 2 in 2,896 cases. Mean value of estimators for the mean value (equal to -1 in general population) of the first component was -0.9957 . Its weight had mean value 0.5007. For the second component of the mixture we had mean value of estimators for the mean value (1 in general population) equals to 0.9957. Its weight was estimated as 0.4993 in the mean.
- III. Again, in the case of location mixture 5,000 samples of volume 1,000 each from the mixture of two Laplace distributions with unite standard deviation and mean values -1 and 1 were simulated. The estimator k^* took true value 2 in 3,041 cases. Mean value of estimators for the mean value (equal to -1 in general population) of the first component was -0.9886 . Its weight had mean value 0.4991. For the second component of the mixture we had mean value of estimators for the mean value (1 in general population) equals to 0.9875. Its weight was estimated as 0.5009 in the mean.

4.2 Financial applications

A. Let us consider some data on stock prices. The data were obtained on the site: <http://finance.yahoo.com/q/hp?a=03&b=12&c=1996&d=02&e=23&f=2016&g=d&s=MSFT>

We analyze stock price changes of GazpromDE for the period from 13-11-2009 till 23-3-2016. Denote by S_j the price of stock at the moment j , and introduce

$$X_j = \log(S_{j+1}/S_j), \quad j = 1, \dots, n,$$

where $n + 1$ corresponds to the number of observed prices. In dataset under consideration it equals to 1,651, so that $n = 1,650$. Applying algorithm of Section 2 we find $k^* = 2$ for a mixture of normal distributions with parameters $(0, 0.01788^2)$ and $(0, 0.301205^2)$. The mixture has weights 0.9948 and

0.005215. The agreement between empirical distribution and this model is not too good (see Figure 1).

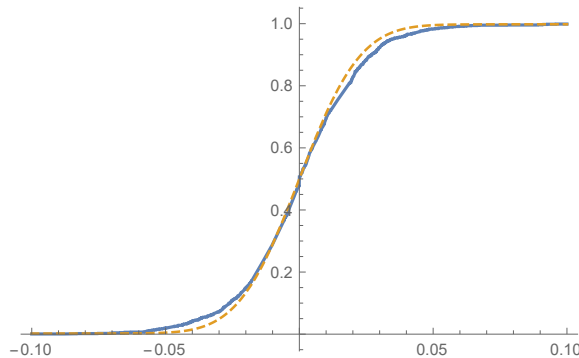


Figure 1: Plot of the empirical distribution function (thick line) and model (dashed line)

Applying maximum likelihood method we find more precise values of normal components variances: 0.00036285 for the first component and 0.1030 for the second. The agreement between empirical data and corrected model is given on Figure 2.

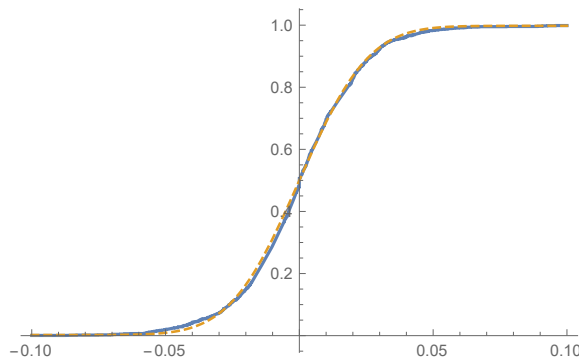


Figure 2: Plot of the empirical distribution function (thick line) and model after maximum likelihood correction (dashed line)

We applied two statistical goodness-of-fit tests to verify agreement between empirical data and corrected model. P-Value for Anderson-Darling test is 0.06708. Cramér-von Mises test has P-Value equal to 0.1476. We see,

that the both test did not reject the hypothesis that sample was drawn from the mixture of normal distributions with zero means.

B. Now we consider exchange rate for US Dollar and Ruble of Russian Federation, the data for period from 11-12-2014 till 17-04-2015 with one minute interval. Denote R_j the exchange rate at moment j . Introduce $\Delta_j = R_{j+1} - R_j$ and remove duplicate values of these differences. Remaining number of observations is $n = 8,180$. After applying the algorithm of Section 2 we find $k^* = 2$ for a mixture of Laplace distributions with zero means and scale parameters 0.1048 and 0.5840. Corresponding weights are 0.8438 and 0.1562. After applying maximum likelihood method we obtain new scale parameters of Laplace distribution. They are 0.1308 and 0.7288. Agreement between the model and empirical data is given on Figure 3.

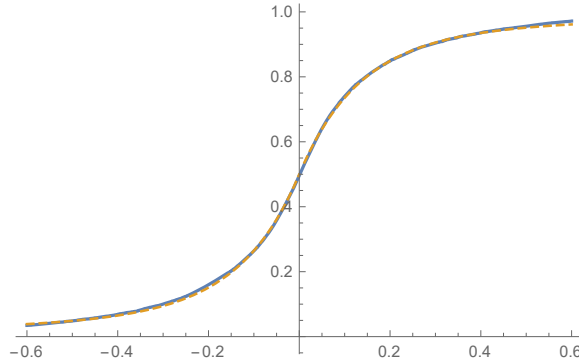


Figure 3: Plot of the empirical distribution function (thick line) and model after maximum likelihood correction (dashed line) for Dollar/RUB exchange rate

We applied two statistical goodness-of-fit tests to verify agreement between empirical data and corrected model. P-Value for Cramér-von Mises test equals to 0.3543, and for Kolmogorov-Smirnov test it is 0.1434. We see, that the both test did not reject the hypothesis that the sample was drawn from the mixture of Laplace distributions with zero means.

4.3 Remarks on multidimensional case

The case when $X = Y \cdot Z$ with multidimensional random vector Z seems to be very interesting. However, one can change it to one-dimensional by taking projection on a direction e , that is by considering inner products

$(X, e) = Y \cdot (X, e)$. Theoretically, it is necessary to consider projections on all vectors e from unit sphere. However, if the vector e is taken “at random”, one often (with probability 1) has this vector at “general position”, and it is enough to use only one projection.

5 Acknowledgment

The work was partially supported by Grant GAČR 16-03708S.

References

- [1] Akhiezer N.I. (1965). The Classical Moment Problem and some related questions in analysis, Oliver & Boyd, 1965 (Translation from Russian Edition, 1961).
- [2] M.G. Krein and A.A. Nudel'man (1977) The Markov moment problem and extremal problems. American Mathematical Society, Providence, Rhode Island, 1977 (Translation from Russian Edition, 1973).